Search Engine

Date
Page

(1) ⇒ Find the list of keywords from all the documents

(2) Now find the TF(t,d) matrix
(t → keyword, d → document)

$$TF(t,d) = \frac{f(t,d)}{\Sigma(t,d)}$$

$f(t,d)$ → No. of occurrence of that keyword in the document

$\Sigma(t,d)$ → total sum of number of keywords present in that document

* [If a keyword comes more than once, we will count it more than once as well]

Eg:- (1) Sun rising sun west
(2) sun rising east east

	sun	rising	east	west
d_1	2/4 = 0.5	1/4 = 0.25	0/4 = 0	1/4 = 0.25
d_2	1/4 = 0.25	1/4 = 0.25	2/4 = 0.5	0/4 = 0
IDF	$\log(\frac{2}{2}) = 0$	$\log(\frac{2}{2}) = 0$	$\log(\frac{2}{1}) = 0.3$	$\log(\frac{2}{1}) = 0.3$

→ Now I have to calculate the importance of each keyword in that corresponding document

Date
Page

(3) Calculate the TF-IDF matrix

→ TF-IDF (t, d, D)

t → keyword, d → document, D → corpus

→ If TF-IDF = 0 for a word in a document, it means that keyword has no importance in defining that document

	sun	rising	east	west
d_1	$0.5 \times 0 = 0$	$0.25 \times 0 = 0$	0	0.075
d_2	$0.25 \times 0 = 0$	$0.25 \times 0 = 0$	0.15	0
IDF	0	0	0.3	0.3

Q) How to calculate the similarity between query string and a document ??

Ans) Take the dot product of their vectors.

Ex:- Query string : sun shining

Q) Calculate the TFIDF matrix for query string

(TF)

	sun	rising	east	west
query	2/3	0/3	0/3	1/3

sun west

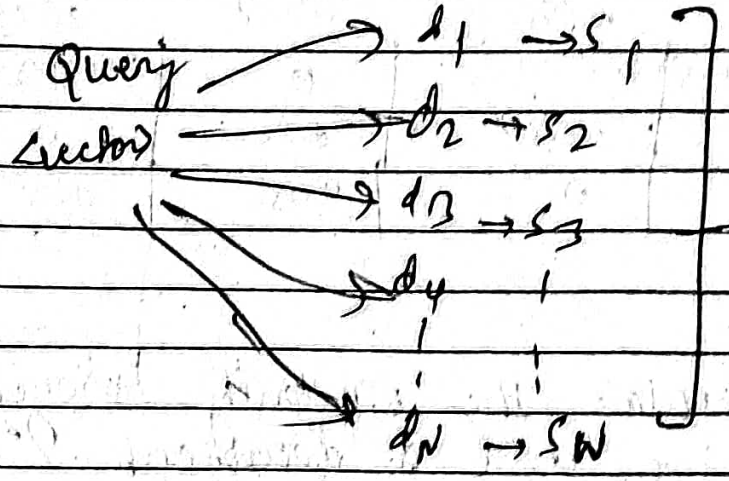
Here as our document has no keyword "shining", so, we ignore it and consider only as a stopword.

TFIDF

	sun	rising	east	west
query	$2/3 \times 0 = 0$	$0/3 \times 0$	$0 \times 0.3 = 0$	$0.3 \times 1 = 0.3$

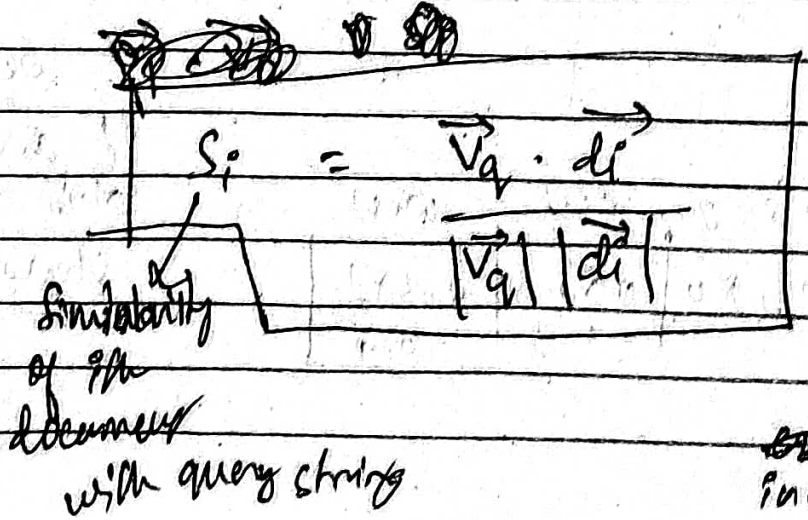
→ Now we have vectors for each document and a vector for the string.

→ We will calculate the similarity values for that query string with each of the documents.



→ The document having higher similarity values are to be placed on top of the search results.

(5) Find the cosine similarities as explained above.



Date
Page

Ex:-

$d_1 = \langle 1, 2, 5, 2 \rangle$

$q = \langle 3, 0, 4, 1 \rangle$

$$\text{sim } S_1 = \frac{1\times3 + 2\times0 + 5\times4 + 2\times1}{\sqrt{1^2+2^2+5^2+2^2}\ \sqrt{3^2+0^2+4^2+1^2}}$$

→ Store each of the vector (TF·IDF) in their own separate (.txt) document